Ultramicroscopy 30 (1989) 299–310
North-Holland, Amsterdam

NORMALIZATION PROCEDURES AND FACTORIAL REPRESENTATIONS FOR CLASSIFICATION OF CORRELATION-ALIGNED IMAGES: A COMPARATIVE STUDY

Michael UNSER *,**

Biomedical Engineering and Instrumentation Branch, National Institutes of Health, Bethesda, Maryland 20892, USA

Benes L. TRUS

Computer Systems Laboratory, Division of Computer Research and Technology, National Institutes of Health, Bethesda, Maryland 20892, USA

and

Alasdair C. STEVEN

Laboratory of Physical Biology, National Institute of Arthritis, Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

Received 21 December 1988; in final form 14 April 1989

We have addressed the problem of optimizing procedures of multivariate statistical analysis (MSA) for identifying homogeneous sets of electron micrographs of biological macromolecules, with a view to averaging over consistent sets of images. Using pre-aligned images of negatively stained protein molecules – known a priori to fall into two subtly different classes – we compared how the capacity to discriminate between them was affected by the normalization procedure used, and by the choice of factorial representation. Specifically, these images were analyzed both after being scaled according to constant minimum and maximum (CMM) values, and after imposing constant values of image mean and variance (CMV). The factorial representations compared were correspondence analysis (CA) and the principal components (PC) formalism. When used with PC, CMM normalization was found to give rise to spurious inter-image fluctuations that were more pronounced than the genuine difference between the two kinds of images; even with CA, CMV proved to be a more satisfactory method of normalization. When CMV was used with CA or PC, both factorial representations yielded qualitatively similar results, although according to a quantitative measure of inter-set discrimination, the performance of PC was slightly superior. Even in the best case, however, the two classes of images – as mapped in factorial space – were not fully resolved. The implications of this observation are discussed with regard to potential ambiguities of image classification in practice.

1. Introduction

The advent of generalized image averaging techniques [1–5] has greatly extended the scope of high resolution electron microscopy of biological macromolecules. However, in order to achieve the highest resolution admitted by a given method of specimen preparation and set of imaging conditions, one must be able to recognize – in the presence of noise – particles that are intrinsically alike. For instance, the particles to be averaged have to lie in the same orientation relative to the beam direction (or substrate plane), and other factors, such as mode of staining, the binding of other molecules, genuine conformational diversity, or artifactual changes sustained during the biochemical isolation procedure or preparation for electron microscopy, may also give rise to heterogeneity.

* Corresponding author.
** Permanent address: INSERM U. 138, Hôpital Henri-Mondor, F-94010 Créteil, France.

To identify homogeneous sets of images in an objective quantitative way, methods of multivariate statistical analysis (MSA), originally developed much earlier and in an entirely different context, have been introduced [6]. According to these procedures, each image is represented as a point in a coordinate system that is derived from the covariance structure of the overall data set (i.e. a factorial representation), and supposedly homogeneous sets or "classes" are defined on the basis of some criterion of clustering or mutual proximity of the points (images) in this space [7].

With a view to defining a procedure that gives optimal inter-class discrimination in practice, we have analyzed a model set of electron micrographs – known to fall into two subtly different classes – in four different ways. We wished to determine how the outcome was affected by two important aspects of the analysis – normalization of the images, and the choice of factorial representation. Since some variability in optical density is to be expected from particle to particle on account of differences in the depth of the stain layer or in photographic development, it is necessary to compensate for this source of (spurious) variability by normalizing each digitized image in the same way. Here we have compared two commonly used normalization conventions: CMM and CMV. Several different forms of factorial representation are used in MSA [8,9] and differ primarily in how the covariance matrix is defined. Of these, CA [10], although a relatively recent and specialized formalism, is the one that has been adopted for applications in electron microscopy [6,7]. Also of interest is the PC formalism [11,9], which has long been used in many applications of MSA, and which possesses the useful property that relative inter-image distances are preserved between real space and factorial space. This latter approach is commonly used in the fields of signal processing and pattern recognition where it is usually referred to as the Karhunen–Loève transform [12,13]. The experimental images were analyzed according to both normalizations and both factorial representations, and the inter-class discrimination evaluated in each case.

The set of images used – taken to be typical of negatively stained data – represent "distal half-fibers" of bacteriophage T7 [14] (see fig. 1). This rod-like molecule ($M_r \sim 90$ kDa) consists of four aligned globules of slightly differing sizes, so that there is a pseudo-dyad at its center. As a result, it is generally difficult to determine which end is which from an unfiltered image of a distal half-fiber along (cf. fig. 2), although this decision may easily be made on the basis of its point of attachment to the "proximal half-fiber" (cf. fig. 1). In our set of distal half-fibers, half were oriented in

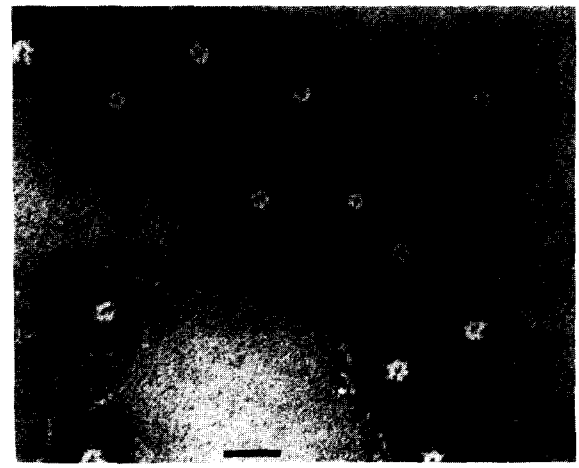

Fig. 1. Purified tail complexes of bacteriophage T7 imaged by bright-field TEM after negative staining with sodium silicotungstate. Bar = 25 nm.

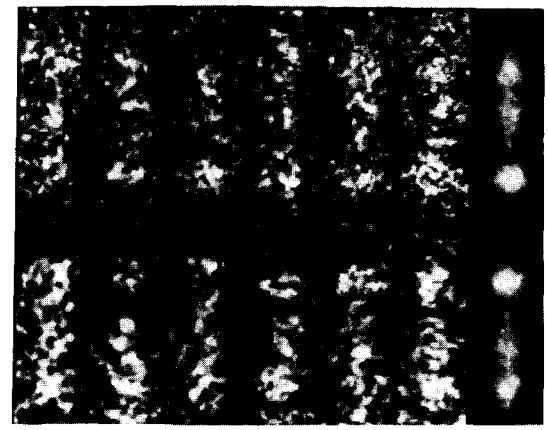

Fig. 2. Examples of individual distal half-fibers in two different orientations. The particles were brought into alignment by correlation techniques. The resolution of these data is 2.0 nm by the SSNR criterion [17]. Each image is 11.9×35.7 nm.

the same way, and half were in the antiparallel configuration.

2. Initial data scaling

2.1. Notations and definitions

The methods described in this paper apply to the processing of a set of N spatially registered images of M pixels each: $\{x_m^{\langle i\rangle}: m=1,\dots,M; i=1,\dots,N\}$. For simplicity, we use a single spatial index m which identifies pixels according to a sequential row-column ordering. This data set may be viewed from two dual points of view depending upon the index on which we focus our attention.

One point of view is to consider the variations from pixel to pixel in a given image ($\langle i\rangle$). In this context, it is useful to define the following image statistics:

- the image first moment or average gray-level value:

$$\bar{x}^{\langle i\rangle} = \frac{1}{M}\sum_{m=1}^{M} x_m^{\langle i\rangle}; \tag{1}$$

- the image second moment or spatial energy (dynamic power):

$$\sigma_x(i)^2 = \frac{1}{M}\sum_{m=1}^{M}\left(x_m^{\langle i\rangle} - \bar{x}^{\langle i\rangle}\right)^2; \tag{2}$$

- minimum and maximum image values:

$$\begin{cases} x_{\min}^{\langle i\rangle} = \min\{x_m^{\langle i\rangle}, & m=1,\dots,M\}, \\ x_{\max}^{\langle i\rangle} = \max\{x_m^{\langle i\rangle}, & m=1,\dots,M\}. \end{cases} \tag{3}$$

The alternative point of view is to consider how a given pixel m varies over the image set: $\{x_m^{\langle i\rangle}, i=1,\dots,N\}$. At this level, we define the mean pixel value

$$\bar{x}_m = \frac{1}{N}\sum_{i=1}^{N} x_m^{\langle i\rangle}, \tag{4}$$

and the local variance:

$$\sigma_{xm}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_m^{\langle i\rangle} - \bar{x}_m\right)^2. \tag{5}$$

Finally, we also define the global average gray level as:

$$\bar{x} = \frac{1}{MN}\sum_{i=1}^{N}\sum_{m=1}^{M} x_m^{\langle i\rangle}. \tag{6}$$

2.2. Normalization procedures

All normalization procedures considered here involve a simple global linear rescaling:

$$z_m^{\langle i\rangle} = a^{\langle i\rangle}\left(x_m^{\langle i\rangle} - b^{\langle i\rangle}\right). \tag{7}$$

The global offset and gain to be applied to each image are specified on the basis of standardizing some property over the entire data set.

For CMM normalization, these properties are the maximum and minimum pixel values and the coefficients are given by:

$$a^{\langle i\rangle} = \frac{z_{\max} - z_{\min}}{x_{\max}^{\langle i\rangle} - x_{\min}^{\langle i\rangle}}, \quad b^{\langle i\rangle} = x_{\min}^{\langle i\rangle} - \frac{z_{\min}}{a^{\langle i\rangle}}, \tag{8}$$

where $z_{\min}$ and $z_{\max}$ are the prescribed values for the extrema of the rescaled images: $\{z_m^{\langle i\rangle}\}$, $i=1,\dots,N$. This approach makes efficient use of available data storage (e.g. one byte per pixel) and is widely used in practice.

For CMV normalization, the coefficients are determined from:

$$a_i = \frac{\sigma_{z0}^2}{\sigma_x^2(i)}, \quad b_i = \bar{x}_i - \frac{\bar{z}_0}{a_i}, \tag{9}$$

where $\bar{z}_0$ and σ_{z0}^2 are prescribed values for the first and second moments of the rescaled images. This normalization is also in common use (e.g., ref. [15]).

3. Factorial representations

In this section, we briefly review the basic PC method. We then provide the relationship with other factorial representations such as PC with mean subtraction, and CA. Finally, we introduce a measure of class separability that will be used later in evaluating discriminatory power.

3.1. Principal components

Images are represented as M-dimensional vectors $\{x_1, \ldots, x_N\}$. We now consider the expansion

$$\hat{x}_i = \sum_{m=1}^{M'} y_m^{\langle i \rangle} u_m, \quad M' \leq \min(M, N) \tag{10}$$

which approximates each image by a weighted sum of some base vectors $\{u_m, m = 1, \ldots, M' \leq N\}$. Principal components have the notable property of providing a representation for which the approximation error for any $M' \leq \min(M, N)$ is minimal, and whose base vectors u are orthogonal [11,9]. The optimal base vectors $\{u_1, \ldots, u_{N'}\}$, or eigen-images, are the eigenvectors of the $M \times M$ scatter matrix: $S_{xx} = XX^T$, where X is the $M \times N$ data matrix $X = [x_1 \ldots x_N]$. The coefficients $\{y_m^{\langle i \rangle}\}$, represented by a $M' \times N$ coefficient matrix $Y = [y_1 \ldots y_N]$, are obtained by simply projecting the data on the sub-space defined by $\{u_m\}$:

$$Y = [y_1 \ldots y_N] = [u_1 \ldots u_{M'}]^T X,$$

and the approximation error is given by:

$$\epsilon^2 = \sum_{i=1}^{N} \| \hat{x}_i - x_i \|^2 = \sum_{i=M'+1}^{\min(M, N)} \lambda_i, \tag{11}$$

where $\{\lambda_i, i = 1, \ldots, N\}$ are the eigenvalues of S_{xx} ordered according to decreasing magnitude. An important property is that the spectrum of eigenvalues $\{\lambda_i\}$ is a measure of the energy distribution across the components of this decomposition.

In our particular case where $M > N$, the rank of S_{xx} is less than or equal to N, implying that there are at most N non-zero eigenvalues $\{\lambda_n, n = 1, \ldots, N\}$. It is therefore computationally advantageous to determine principal components by diagonalizing the $N \times N$ inner product matrix: $R_{xx} = X^T X$. This procedure uses the property that the non-zero eigenvalues of (XX^T) and $(X^T X)$ are identical and that the corresponding eigenvectors of these matrices ($\{u_n\}$ and $\{v_n\}$, respectively), are related by the following equation:

$$u_n = X \cdot v_n / \sqrt{\lambda_n} \quad (n = 1, \ldots, N'). \tag{12}$$

In addition, we can show that the optimal expansion coefficients in eq. (10) may be simply determined by

$$Y = [y_1 \ldots y_N] = \left[\sqrt{\lambda_1}\, v_1 \ldots \sqrt{\lambda_{N'}}\, v_{N'}\right]^T. \tag{13}$$

In summary, the computational steps involved in the determination of PC for a given set of electron micrographs are the following:

(i) Perform the appropriate initial data scaling.

(ii) Compute the $N \times N$ inner product matrix $R_{xx} = X^T X$ which amounts to evaluating the cross-correlations between all possible pairs of images. This is usually by far the most computationally intensive task.

(iii) Determine its eigenvectors and eigenvalues $\{(v_i, \lambda_i), i = 1, \ldots, N\}$.

(iv) Compute the coordinates in factorial space according to eq. (13).

(v) Eventually compute the eigen-images $\{u_n\}$ using eq. (12).

The data are then analyzed in terms of the coordinates (or components) in this reduced projection space. The number of significant components is usually determined empirically, based on the magnitudes of the eigenvalues [9]. The other components are essentially due to noise and are disregarded for the purposes of classification or data reduction.

3.2. Relationship with other factorial representations

Almost any other factorial representation can be obtained using PC, provided that the data has been previously normalized in an appropriate fashion. The general equation for such a normalization is

$$z_m^{\langle i \rangle} = a_m^{\langle i \rangle} \left(x_m^{\langle i \rangle} - b_m^{\langle i \rangle} \right), \tag{14}$$

which is very similar to eq. (7) except that a and b are now allowed to vary as functions of the spatial index m.

Often one is primarily interested in the differences between measurements, in which case it is natural to perform the analysis on variables that are centered about their mean. The appropriate normalization coefficient for PC with mean subtraction are as follows:

$$a_m^{\langle i \rangle} = a_0, \quad b_m^{\langle i \rangle} = \bar{x}_m, \tag{15}$$

where a_0 is an arbitrary positive constant.

Another popular method is CA [10,6]. By design, this method requires the measurement values to be positive. This constraint has to be taken care of in the case where the data is initially scaled using either min/max and mean/var scaling techniques. The equivalence between CA and a normalized form of PC is well known [16]. Translating this basic result (ref. [16], pp. 282 and 311) into our notation, it is relatively straightforward to show that CA is equivalent to PC with the following normalization equations:

$$a_m^{\langle i\rangle} = \frac{1}{\sqrt{MN\bar{x}_m\bar{x}^{\langle i\rangle}}}, \quad b_m^{\langle i\rangle} = \frac{\bar{x}_m\bar{x}^{\langle i\rangle}}{\bar{x}}. \tag{16}$$

We note that when the images have been initially scaled using CMV, $\bar{x}^{\langle i\rangle}$ is equal to $\bar{x}$ and the normalization is equivalent to adjusting each pixel value by subtracting $\bar{x}_m$ and then dividing by $\sqrt{\bar{x}_m}$.

3.3. A measure of class separability

In the context of correlation averaging, the usefulness of a given factorial representation resides in its capacity to distinguish differences between subsets of particles. At this stage of the analysis, a given image i is represented by a reduced set of coordinates: $\{y_m^{\langle i\rangle},\ m = 1,\ldots, M' \le N\}$. In the event that the particles are assigned to K distinct classes with n_k $(k = 1,\ldots, K)$ particles respectively, the separation power or discriminability of a given component $y_m^{\langle i\rangle}$ is given by the ratio between the inter-class variance and the intra-class variance:

$$\beta_{ym} = \frac{\sum_{k=1}^{K} n_k\left[\bar{y}_m(k) - \bar{y}_m\right]^2}{\sum_{k=1}^{K} n_k\sigma_{ym}^2(k)}. \tag{17}$$

In this equation, $\bar{y}_m(k)$ and $\sigma_{ym}^2(k)$ denote the mean and variance of $y_m^{\langle i\rangle}$ in class k. These values are determined from equations similar to (1) and (2) with the summation being performed over the corresponding subset of n_k particles. $\bar{y}_m$ is the mean value over the entire set.

The components with the most potential for classification or clustering are those that give the largest values of β_{ym}. On the other hand, the components with values of β_{ym} close to zero may be regarded as noise. We note that this separability criterion relies on the use of distance (in the Euclidean sense) to measure similarity between particles.

4. Results

Intact tail complexes of bacteriophage T7 have six kinked fibers (fig. 1), whose distal parts are ~16 nm long by ~4 nm wide. They have a nodular substructure with a definite polarity that is evident in averaged images, but much less so prior to averaging [14] (cf. fig. 2). Sets of distal half-fibers in both orientations, mutually aligned by correlation methods, provided a set of model data with the attributes of real experimental images.

This data set was analyzed in four ways (two factorial representations, and two normalization procedures). The four corresponding eigenvalue spectra were calculated (fig. 3 and table 1). According to the criterion that significant factors are those whose eigenvalues are sufficiently large to rise above the background continuum, one of the analyses (PC + CMM) appears to have two

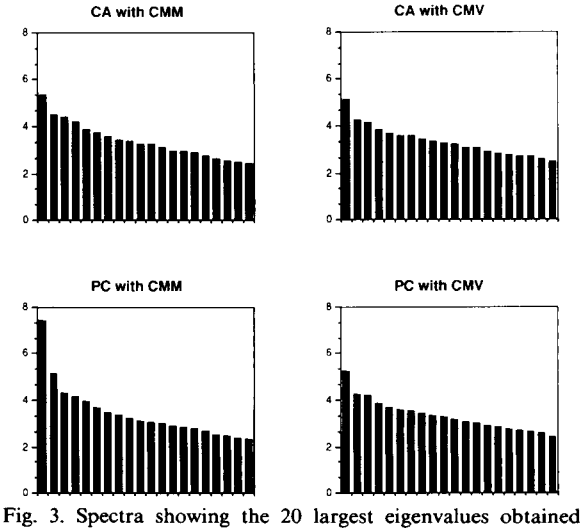

Fig. 3. Spectra showing the 20 largest eigenvalues obtained when the set of experimental distal half-fiber images was analyzed according to the four methods listed.

Table 1
Eigenvalues (λ_i) (expressed as percentages of the total energy) and inter-class discrimination indices (β_i) calculated for the first four factors corresponding to each of the four different multivariate statistical analyses; the values of β that are significantly above zero are underlined; the β's evaluated for the remaining factors were all below 5×10^{-2}

Methods	λ_1, β_1	λ_1, β_2	λ_3, β_3	λ_4, β_4
CMM and CA	5.34%, 1.62	4.50%, 0.05	4.43%, 0.03	4.22%, 0.00
CMV and CA	5.13%, 1.99	4.26%, 0.07	4.17%, 0.03	3.83%, 0.02
CMM and PC	7.43%, 0.10	5.16%, 1.44	4.30%, 0.00	4.16%, 0.05
CMV and PC	5.26%, 2.19	4.26%, 0.10	4.23%, 0.00	3.85%, 0.01

significant factors, whereas the other three have only one.

For each factor of each MSA formalism, its capacity to discriminate between the two subsets of images was evaluated in terms of the parameter β (see section 3.3, eq. (17)). The factors with the greatest discriminatory power (and hence the greatest potential for classification and clustering) are those that give the largest values for β, whereas those whose β-values are close to zero may be regarded as noise. In this capacity, β closely resembles a signal-to-noise ratio. The resulting

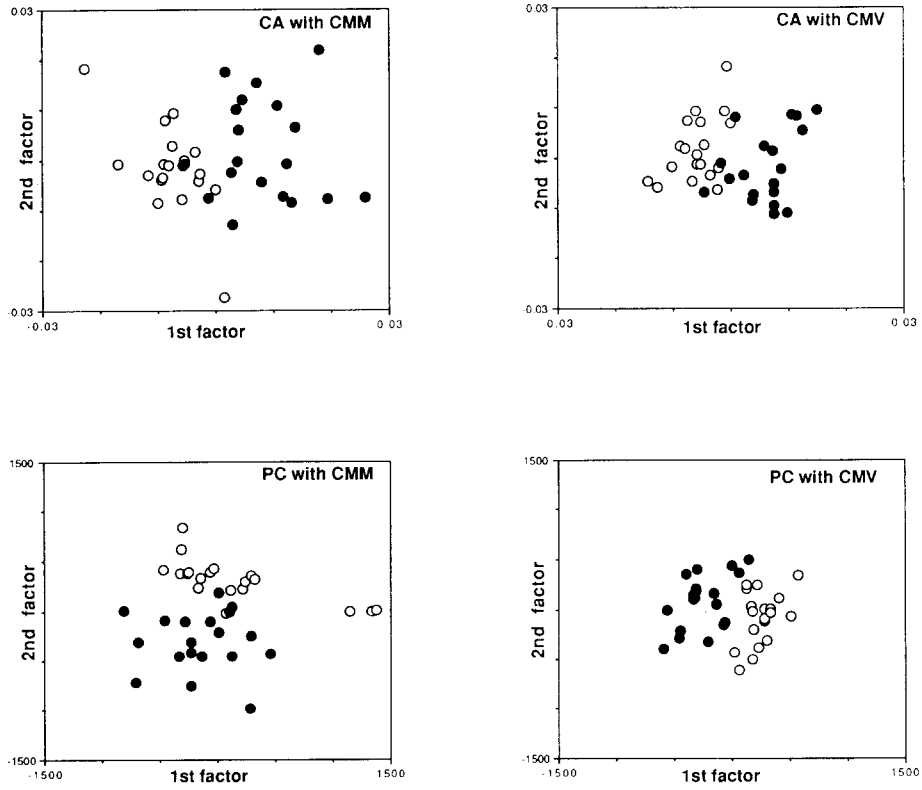

Fig. 4. Factorial maps associated with the two most significant factors obtained when each of four different MSA were applied to a set of 38 distal half-fiber images of which half were oriented in parallel (○) and half in antiparallel (●) (cf. fig. 2). All factors obtained using a particular factorial analysis method (CA or PC) are shown on the same scale. These graphs illustrate the property that the use of CMV tends to reduce the intra-class scatter when compared to CMM.

values are listed in table 1. In each case, only one factor detects a significant distinction between the two subsets: with (PC + CMV), (CA + CMM), and (CA + CMV) it is the first factor, but with (PC + CMM), it is the second (i.e., in this MSA the factor with the highest eigenvalue is, in fact, spurious).

Two-factor plots are also presented for each analysis (fig. 4), and they convey the same effect as the tabulated values of β. With (PC + CMM), the images are randomized in the first factor (horizontal axis) and segregated in the second (vertical axis), whereas the situation is reversed for the other three mappings. Therefore, when CMM normalization is used in conjunction with PC, spurious fluctuations in the pixel densities constitute a more important source of variation than the genuine structural difference between the two kinds of images. Nevertheless, when PC is used with CMV normalization, the best margin of discrimination, i.e. the highest β-value, is obtained.

It is of interest to compare the first few eigen-images of the respective analyses (fig. 5). In a grey-scale representation of an eigen-image (as shown here), areas that are very dark or very light denote locations where high levels of variability occur for the factor in question, whereas intermediate grey tones denote relatively small fluctuations among the images. The first few eigenimages are recognizably the same in each case, except for an offset by one image with (PC + CMM), whose leading eigenimage relates to the spurious normalization-related effect described above. Thus, with this set of negatively stained data at least, essentially the same characteristics of primary variation are picked up in each of the four analyses.

Finally, we note that, even with the 2-factor plot that corresponds to the largest margin of discrimination achieved in these experiments (PC + CMV, fig. 4), the two classes of images segregate but are not fully resolved from one another. Thus, if we had not already been aware of their division into two classes, it would have been a ticklish business to distinguish between them from this plot. The implications of this observation for the practical problem of assigning experimental data into coherent classes [7] are discussed further below.

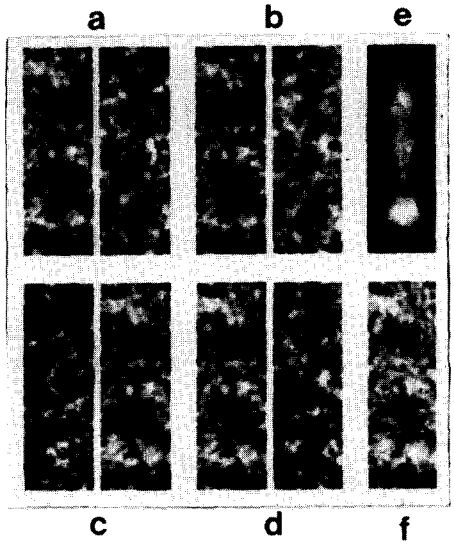

Fig. 5. Factorial analysis of T7 distal half-fibers: (a) first two eigen-images (CA + CMM); (b) first two eigen-images (CA + CMV); (c) first two eigen-images (PC + CMM); (d) first two eigen-images (PC + CMV); (e) image average; (f) difference between average images in alternative (up and down) orientations.

5. Discussion

5.1. A spurious factor occurs when CMM normalization is used in conjunction with principal components

The most notable difference among the four analyses was the spurious factor encountered with (PC + CMM). The number of significant factors in a given MSA is usually decided upon by referring to a discontinuity in the slope of the eigenvalue spectrum [18]. By this criterion, the spectra in fig. 3 suggest a single significant factor for CA with both normalizations and for (PC + CMV). However, with (PC + CMM), the spectrum suggests two significant factors, and it is noteworthy that the spurious first eigenvalue exceeds the genuinely significant second one by a substantial margin. The origin of this spurious factor is that CMM normalization is subject to the vagaries of single pixels i.e. the extreme densities. (It could,

for instance, be offset in a major way by a single dust particle on a scanned negative, although such was not the case with these experiments.) Whether or not these fluctuations turn out to be the dominant factor will depend on how pronounced the real differences are. Notwithstanding the present experiments make the point that CMM-related fluctuations can dominate a genuine, albeit subtle, source of variation. CMV normalization, on the other hand, depends on the entire distribution of pixel densities, and accordingly, is a stabler procedure.

It is noteworthy that this source of spurious variation does not occur with CA, because CA effects an implicit post-normalization that is more-or-less equivalent to a spatial mean standardization (cf. eq. (16)). In any case, better inter-class discrimination was achieved with CA when CMV normalization was used, and we conclude that this normalization is to be preferred over CMM with either factorial representation.

5.2. *Which factorial representation gives best inter-class discrimination in practice?*

In principle, it would be possible to undertake image classification on real-space representations. However, a major problem with doing so is that the high noise levels of non-averaged micrographs would tend to swamp the fine details on which a valid classification would ultimately depend. The advantage of a factorial representation for this purpose is to greatly reduce noise levels * while preserving most of the inter-class differences. In practice, however, the residual levels of noise can be substantial relative to the mean differences as indicated by the scatter of the points in fig. 4, as also observed in many previous applications of CA (e.g., refs. [6,17]). This being the case, a question of practical importance is: which of the many possible factorial representations most strongly accentuates inter-class differences, and therefore has the greatest potential for accurate inter-class discrimination?

The essential difference between CA and PC lies in their respective scaling factors which lead to a different concept of inter-image distances in factorial space. The underlying metric in PC space is the conventional Euclidean distance. According to classical decision theory [13], this is the metric of choice for classifying objects corrupted by additive white Gaussian noise, which makes PC preferable in such cases. In contrast, CA expresses proximity between elements in terms of chi-square distances, and so puts comparatively more weight on low signal values, which may or may not be a desirable property. Finally, we note that the factors in PC are invariant with respect to global additive and multiplicative scaling of the data, whereas in CA this invariance property holds only for multiplicative scaling.

With the present data – experimental micrographs of negatively stained protein molecules – the global S:N was estimated to be 0.61*, although the specifics of the noise distribution are not known, and are unlikely to conform precisely to some idealized distribution. To a first approximation (and once the spurious factor of (PC + CMM) was discounted), all four analyses gave similar results – a single significant factor (fig. 3), the same unit cell locations identified as particularly variable (fig. 5), and a segregation of the two classes along the factorial axis (fig. 4). However, in quantitative terms (table 1), (PC + CMV) gave the best discrimination with a 10% higher value of parameter β than was realized with (CA + CMV).

5.3. *Is there an optimal factorial representation?*

The ideal factorial representation would maximize class discriminability. As shown in the appendix, a precise knowledge of the statistics (mean and covariance) of underlying image classes is needed to define such a representation. In practice, however, there is no way to obtain this information, and we can at best approximate this hypothetical solution by using procedures such as PC and CA. Nevertheless, there are some useful

* For additive white Gaussian noise, it can be demonstrated using perturbation methods that the total noise energy tends to be evenly distributed over all dimensions of factorial space.

* Square root of the ratio of the dynamic power of the mean image to the average power of the residual (i.e. original-mean) images.

points to be learned from studying the properties of such a solution. First, as discussed in the appendix, the optimal solution is closest to PC when the data are standardized so that the intra-class noise variance is the same for all pixels. Thus, it is important to employ preprocessing or scaling procedures that try to achieve this goal, and the use of CMV normalization is a good step in this direction. Second, there is the issue of the number of significant factors, which should not be greater than the number of distinct classes minus one. In particular, for a two-class problem such as the one studied here, there should be only one significant factor which corresponds to the difference between the respective mean images. It is evident from comparing fig. 5f and the eigen-images in figs. 5a to 5d that this factor is reasonably well extracted by each procedure.

5.4. Implications for classification of experimental electron micrographs

One point of concern raised by these experiments has to do with the feasibility of performing reliable classification in practice. The level of discrimination achieved between two genuinely different classes of images was, in the best case (fig. 4d), a segregation in factorial space and not full resolution into two separate clusters. That is to say, it is not clear how successful formal classification algorithms [7] are likely to be when applied to statistically comparable data in a practical setting. With these data, a reasonably successful outcome (92%) would be obtained by simply defining the dividing-point between two classes as the zero-value of the factorial coordinate in the last plot in fig. 4. However, this strategy presupposes that we know that there are two classes and that they have equal numbers of members, premises that are not likely to be generally met. We have also applied the KMEANS [19] and several hierarchical clustering algorithms [7,19] to the data as represented by the two-factor coordinates from the (PC + CMV) analysis (the most avantageous), again in the rather artificial situation of foreknowledge that the images fall into two classes (see table 2). These experiments confirm what is already suggested by visual inspection of the two-factor plots (fig. 4), that if the noise level is sufficiently high, foolproof classification is not to be achieved even with the use of quantitative algorithms. The use of a larger number of factors did not improve these results, since those components contain no discriminating information (their β's are all close to zero, cf. table 1) and their only effect is to increase the level of noise.

Table 2
Classification of T7 distal half-fiber images with several clustering algorithms in a two factor space (PC + CMV); we used standard procedures available in the SYSTAT statistical package [20]; the KMEANS algorithm is according to ref. [19]; the various hierarchical algorithms correspond to different linkage criteria [13,19]; the specified options were linkage = centroid, median, single and complete, respectively; the centroid method is equivalent to the hierarchical ascendant clustering algorithm described in ref. [7]; these results are predicated on the assumption that the data fall into two classes, which is an artificially advantageous situation, since this information is not generally available in practice; a classification error corresponds to assigning a given particle to the wrong class

Method	Set A	Set B	Errors
Ideal classification	(19 up, 0 down)	(0 up, 19 down)	0%
Random assignment	(–, –)	(–, –)	50%
Iterative:			
KMEANS	(17 up, 3 down)	(2 up, 16 down)	13.2%
Hierarchical:			
Centroid [a]	(15 up, 3 down)	(3 up, 16 down)	18.4%
Median	(13 up, 3 down)	(6 up, 16 down)	23.7%
Nearest neighbor	(19 up, 3 down)	(0 up, 16 down)	7.9%
Farthest neighbor	(13 up, 12 down)	(6 up, 7 down)	47.4%

[a] The centroid linkage method initially isolated a set containing only one image.

In such marginal situations, any incremental improvement in the *initial* S : N ratio, achieved either by improvements in specimen preparation or imaging techniques, or by refinements of the procedures for aligning the particles, or in performing the MSA, is to be welcomed. Thus the slight margin of superiority in the performance of (PC + CMV) over (CA + CMV) observed here may be of greater practical significance than appears at first sight. In this context, we have also tried some other methods of pre-processing the data that might reduce noise. By windowing the particles with a binary template that eliminates those back-

ground pixels outside the particle's contour (50%) [18], the largest β obtained was 2.42 for (PC + CMV). We have also tried low-band-pass filtration with a bell-shaped filter falling to zero at (1.5 nm)$^{-1}$, based on the premise that any components beyond this spatial frequency must be noise. However, no significant improvement in inter-class separation was recorded in this case. It might also be that noise in strong low-frequency components of the image may obscure classification distinctions based on relatively fine details. To test this, we applied a mid-band-pass filter [1/7.5 to 1/2 nm^{-1}] falling to zero at frequencies below 1/8 and above 1/1/5 nm^{-1}. This procedure resulted in a β-value of 2.37 for (PC + CMV), which, as with masking, indicates a small improvement. However, in neither case was the slight improvement achieved sufficient to resolve the two classes fully in factorial space.

Classification of images in a space where they are not fully resolved is analogous to the proposition of "super-resolution" in optics. As an attempt to tackle this problem, one may apply clustering algorithms (e.g., ref. [7]), but as the results given above (table 2) indicate, they are not foolproof, particularly for subtle structural distinctions in the presence of substantial noise levels. Another pragmatic approach is to concede some degree of objectivity, and perform "supervised classification" (work in progress). In this approach which bears some relation to the KMEANS or ISODATA clustering algorithms [8,13,19], two or more archetypal particles are chosen, and the remaining images assigned according to their respective proximities (in factorial space) to these references (or to the current inclass averages). This procedure allows for elimination of those images which are not closer to one reference than the others by some pre-specified margin. The reliability of such analyses may be assessed in terms of its stability with respect to variations in choice of reference particles, and to the selections made by different observers.

6. Conclusions

With the goal of defining methods of multivariate statistical analysis that are optimal for use in correlation averaging, we find that:

(1) The use of CMM (constant minimum and maximum) normalization introduces a substantial amount of scaling-associated noise, and CMV (constant mean and variance) normalization is definitely to be preferred.

(2) The most appropriate choice of factorial representation depends on the noise statistics. With a set of micrographs of negatively stained protein molecules (experimental data, whose noise statistics are not well defined), correspondence analysis and principal components gave rather similar results, suggesting that both procedures are comparably effective with data of this kind. However, principal components did achieve a slightly better degree of discrimination between the two classes of images present.

(3) Even with the best procedure, these two classes were not fully resolved in factorial space. The implication is that problematic levels of noise persist even when the images are mapped into the most advantageous factorial space. Such noise levels are sufficiently high to compromise the reliability of classifications that depend on fine details, although such distinctions must be recognized if the goals of correlation averaging are to be fully realized. This appears to be a fundamental limitation which, unless some solution can be found, may severely circumscribe the prospects for generally applicable methods of classification and correlation averaging as applied to electron micrographs of biological macromolecules.

Acknowledgements

We are grateful to Dr. J.V. Maizel (Advanced Scientific Computing Laboratory, Frederick, MD, USA) and F.W. Studier (Brookhaven National Laboratory, Upton, NY, USA) for making available the micrographs used in this study.

Appendix. Factorial representation for maximal class discrimination

In this appendix, we consider the problem of defining a factorial representation that is optimal in the sense that it maximizes class separation. More specifically, we are seeking a set of compo-

nents: $\{ y_m = \boldsymbol{u}_m^{\mathrm{T}} \boldsymbol{x},\ m = 1, \ldots, M' \}$ of minimal dimension that maximizes the ratios defined by eq. (17). For this purpose, we assume that there are K distinct classes $\{ \omega_k,\ k = 1, \ldots, K \}$. Each class ω_k is characterized by an a priori probability $P(\omega_k)$ (the overall fraction of images falling into this particular category), a mean vector $E\{ \boldsymbol{x} \mid \omega_k \}$ (which corresponds to the average image within this group), and an $M \times M$ covariance matrix $\boldsymbol{C}_k$ that specifies the intra-class variability of the measurements relative to the class mean. We denote by $E\{ \boldsymbol{x} \}$ the expected global average image which is given by:

$$E\{ \boldsymbol{x} \} = \sum_{k=1}^{K} P(\omega_k) E\{ \boldsymbol{x} \mid \omega_k \}. \tag{A.1}$$

The basis vectors defining the optimal transformation are found by maximizing the measure of discriminability:

$$\beta_y = \sigma_{\mathrm{b}}^2 / \sigma_{\mathrm{w}}^2, \tag{A.2}$$

where

$$\sigma_{\mathrm{b}}^2 = \sum_{k=1}^{K} P(\omega_k) \left[E\{ y \mid \omega_k \} - E\{ y \} \right]^2 \tag{A.3}$$

$$\sigma_{\mathrm{w}}^2 = \sum_{k=1}^{K} P(\omega_K) E\left(\left[y - E\{ y \mid \omega_k \} \right]^2 \mid \omega_k \right), \tag{A.4}$$

are the expected inter-class and intra-class variances of the factor y, respectively. Since $y = \boldsymbol{u}^{\mathrm{T}} \boldsymbol{x}$, this ratio can be shown to be equal to:

$$\beta_y = (\boldsymbol{u}^{\mathrm{T}} \boldsymbol{B} \boldsymbol{u}) / (\boldsymbol{u}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{u}), \tag{A.5}$$

where $\boldsymbol{B}$ is the $M \times M$ expected inter-class scatter matrix that characterizes the differences between the group means:

$$\boldsymbol{B} = \sum_{k=1}^{K} P(\omega_k) \left[E\{ \boldsymbol{x} \mid \omega_k \} - E\{ \boldsymbol{x} \} \right] \times \left[E\{ \boldsymbol{x} \mid \omega_k \} - E\{ \boldsymbol{x} \} \right]^{\mathrm{T}}, \tag{A.6}$$

and where $\boldsymbol{W}$ is the $M \times M$ expected intra-class scatter matrix:

$$\boldsymbol{W} = \sum_{k=1}^{K} P(\omega_k) E\left(\left[\boldsymbol{x} - E\{ \boldsymbol{x} \mid \omega_k \} \right] \times \left[\boldsymbol{x} - E\{ \boldsymbol{x} \mid \omega_k \} \right]^{\mathrm{T}} \mid \omega_k \right) = \sum_{k=1}^{K} P(\omega_k) \boldsymbol{C}_k, \tag{A.7}$$

which also corresponds to the weighted average of the individual covariance matrices. By differentiating (A.5) with respect to $\boldsymbol{u}$, it follows that the set of optimal basis vectors are obtained from the generalized eigen-solutions of the characteristic equation

$$\boldsymbol{B}\boldsymbol{u} = \beta \boldsymbol{W}\boldsymbol{u}. \tag{A.8}$$

This result can be related to the classical theory of discriminant analysis (e.g., ref. [13]). The interpretation of this equation is simplified if we express it in the space of the transformed variables $\boldsymbol{x}' = \boldsymbol{W}^{-1/2} \boldsymbol{x}$ with corresponding inter-class scatter matrix $\boldsymbol{B}' = (\boldsymbol{W}^{-1/2})^{\mathrm{T}} \boldsymbol{B} (\boldsymbol{W}^{-1/2})$ and an identity intra-class covariance matrix. When $\boldsymbol{W}$ is a diagonal matrix (which is certainly true for independent noise), this change of coordinate system is obtained by simple standardization, that is, by dividing the value of each pixel by its corresponding noise standard deviation. By making this change of variable and defining $y = \boldsymbol{u}'^{\mathrm{T}} \boldsymbol{x}'$, we find that:

$$\boldsymbol{B}'\boldsymbol{u}' = \beta \boldsymbol{u}', \tag{A.9}$$

which is an expression that is very similar to the one defining the axes of the principal components representation (e.g., ref. [9]). This indicates that the optimal solution corresponds to a projection of the standardized data on the principal axes of the corresponding inter-class scatter (or mean difference) matrix. In other words, the optimal representation is derived from the principal components of the standardized mean vectors. Based on this theoretical result, we can make the following observations:

(i) From eqs. (A.6) and (A.1), we know that the rank of the matrix $\boldsymbol{B}$ (or equivalently $\boldsymbol{B}'$) is at most equal to $K - 1$, where K is the total number of classes. This implies that there are at most $K - 1$ optimal factors with non-zero β-values. The expected values for β in all directions of the feature space that are perpendicular to this subspace are zero.

(ii) The simplest case occurs when the matrix $\boldsymbol{W}$ is proportional to the identity matrix. This condition is clearly satisfied when the particles are corrupted with additive white noise. In this case, the optimal factorial representation is simply a

projection of the data on the sub-space of the group means. In particular, for a two class classification problem, there is only one optimal factor whose corresponding eigen-image, assuming that $P(\omega_1) = P(\omega_2) = 1/2$, is proportional to the difference of the group means, that is: $\boldsymbol{u} = E\{\boldsymbol{x} \mid \omega_1\} - E\{\boldsymbol{x} \mid \omega_2\}$.

References

[1] W. Baumeister and W. Vogell, Eds., Electron Microscopy at Molecular Dimensions (Springer, Berlin, 1979).

[2] J. Frank, in: Computer Processing of Electron Microscopic Images, Ed. P.W. Hawkes, (Springer, Berlin, 1980) p. 188.

[3] W.O. Saxton and W. Baumeister, J. Microscopy 127 (1982) 127.

[4] J. Frank, Ultramicroscopy 1 (1975) 159.

[5] J. Frank, A. Verschoor and M. Boublik, Science (USA) 214 (1981) 1353.

[6] M. van Heel and J. Frank, Ultramicroscopy 6 (1981) 187.

[7] M. van Heel, Ultramicroscopy 13 (1984) 165.

[8] K.V. Mardia, J.T. Kent and J.M. Bibby, Multivariate Analysis (Academic Press, London, 1979).

[9] I.T. Jolliffe, Principal Component Analysis (Springer, Berlin, 1986).

[10] J.-P. Benzécri, L'Analyse des Données, Vol. 1: La Taxionomie (Dunod, Paris, 1979).

[11] H. Hotteling, J. Educ. Psychol. 24 (1933) 417.

[12] S. Watanabe, in: Trans. 4th Conf. on Information Theory, Prague, 1965, p. 635;
S. Watanabe, in: Pattern Recognition: Introduction and Foundation, Ed. J. Sklansky (Dowden, Hutchinson and Ross, 1973) p. 146.

[13] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis (Wiley, New York, 1973).

[14] A.C. Steven, B.L. Trus, J.V. Maizel, M. Unser, D.A.D. Parry, J.S. Wall, J.F. Hainfeld and F.W. Studier, J. Mol. Biol. 200 (1988) 351.

[15] J.L. Carrascosa and A.C. Steven, Micron 9 (1978) 199.

[16] L. Lebart, A. Morineau and J.-P. Fénélon, Traitements des Données Statistiques (Dunod, Paris, 1979).

[17] M. Unser, B.L. Trus and A.C. Steven, Ultramicroscopy 23 (1987) 39.

[18] J. Frank, A. Verschoor and M. Boublik, J. Mol. Biol. 161 (1982) 107.

[19] J.A. Hartigan, Clustering Algorithms (Wiley, New York, 1975).

[20] L. Wilkinson, SYSTAT: The System for Statistics (SYSTAT, Evanston, IL, 1987).